International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN(P): 2249-6890; ISSN(E): 2249-8001
Vol. 7, Issue 3, Jun 2017, 105-112

NOISE REDUCTION BASED ON ROBUST SPEECH AND NON-SPEECH DETECTION IN VEHICULAR ENVIRONMENTS

JEONG-SIK PARK[1], GIL-JIN JANG[2] & YONG-HO SEO[3]

[1]*Department of Information and Communication Engineering, Yeungnam University, Republic of Korea*

[2]*School of Electronics Engineering, Kyungpook National University, Republic of Korea*

[3]*Department of Intelligent Robot Engineering, Mokwon University, Republic of Korea*

ABSTRACT

Background/Objectives

A variety kinds of noise reduction approaches have achieved different performance according to noise types and user environments. Therefore, a suitable noise reduction approach needs to be investigated in consideration of a specific user environment.

Methods/Statistical Analysis

Most noise reduction methods are affected by the correctness of speech and non-speech detection, because a primary process of the methods is to estimate noise components from non-speech regions. In this study, we propose an efficient noise reduction approach to be adopted in vehicular environments. The proposed approach is based on robust speech and non-speech detection approach based on variance of spectral energy in frequency bin.

Findings

To observe the property of vehicular noise signals sophisticatedly, we recorded a set of noise sounds inside a car while driving the car. Vehicular noise signals have a general property of stationary noise types, indicating slowly changing signal characteristics. The property is observed more significantly in frequency domain compared to time domain. This observation concludes that in vehicular environments, a method of reducing stationary noise signals is quite applicable. The proposed approach considers the general tendency of spectral energy that speech regions indicate higher spectral variance than non-speech regions do. Thus, the method can improve the performance of noise reduction. Once a non-speech region is detected according to our mechanism, the spectral energy is regarded as the energy of noise components. So, the estimated spectral energy is used for suppressing noise components in subsequent speech regions.

Improvements/Applications

We evaluated the efficiency of the proposed approach via spectral analysis. The proposed voice activity detection method demonstrated superior noise reduction performance compared to the conventional method in terms of SNR.

KEYWORDS: Noise Reduction; Vehicular Environment; Speech and Non-Speech Detection; Spectral Analysis; Spectral Energy & Noise Components

Received: Mar 27, 2017; **Accepted:** Apr 24, 2017; **Published:** May 05, 2017; **Paper Id.:** IJMPERDJUN201713

Original Article

1. INTRODUCTION

Nowadays, people possess a variety kinds of speech-related devices such as smartphones, and uses speech applications like speech communication and speech recognition[1,2]. As electronic devices have penetrated home and vehicular environments, more variety kinds of speech applications have been used. However, such applications may be interfered by adverse conditions in outdoor environments[3,4]. Background noise is the most representative interference[5].

Tremendous efforts have been made to reduce various types of noise signals for the purpose of improving speech quality[6-9]. A general way of noise reduction firstly estimates the noise signals in background regions having no speech signals and removes them in speech regions[10-12]. Many techniques such as spectral subtraction and Wiener filtering have been successfully employed.

Although various noise reduction methods have been applied for many speech-related applications, the approaches reported different performance according to noise types and user environments. For this reason, a suitable noise reduction approach should be investigated in consideration of a specific user environment. This study proposes an efficient approach with regard to vehicular environments.

This paper is organized as follows. Section 2 explains the conventional noise reduction approaches. The proposed method is introduced in Section 3, and the experiments and results are discussed in Section 4. And, Section 5 concludes this study.

2. THE CONVENTIONAL NOISE REDUCTION METHODS

Most of the conventional noise reduction methods have been introduced for improving automatic speech recognition performance. According to the approaches, noise signals are usually assumed to be stationary, additive and uncorrelated to speech. The common procedures consist of estimation of the noise components and elimination of them in noisy speech. If the background noise is evolving more slowly than the speech, i.e., if the noise is more stationary than the speech, it is easy to estimate the noise components during pauses in speech. Finding the pauses in speech signals is based on checking how close the estimate of the background noise is to the signal in the current frame. The Voice Activity Detection (VAD) techniques are generally used for this work. The VAD gives values of zero and one as indicators of the voice activity in each frame and enables to update the estimate of the background noise spectrum on the frames that have zero value, using the following formula.

$$\left|N(\omega,n)\right|^2 = \lambda \cdot \left|N(\omega,n-1)\right|^2 + (1-\lambda)\cdot\left|X(\omega,n)\right|^2 \quad (1)$$

where $X(\omega,n)$ and $N(\omega,n)$ are the spectrum of the noisy speech and noise signals, respectively. λ and n refer to a decay rate coefficient flattening the spectrum and the index of the current frame, respectively. On consecutive noise frames, this equation becomes true and the VAD indicates a value of zero.

There are also more sophisticated methods for estimating the noise components. The most representative method is the spectral subtraction[13-15]. This technique assumes that the noise and speech are uncorrelated and additive in the time domain. In that case, the power spectrum of the noisy speech signal is the sum of the noise and the speech spectra. The method also assumes that the noise characteristics change slowly relative to those of speech signals, so that the noise spectrum estimated on non-speech frames can be used for suppressing the noise components contaminating the speech. Let

$x(t)$, $s(t)$ and $n(t)$ be the noisy speech signal, original clean speech signal, and additive noise signal, respectively. According to spectral subtraction approaches, a clean speech spectrum ($\left|\hat{S}(\omega)\right|$) can be estimated by subtracting an average noise spectrum ($\left|\hat{N}(\omega)\right|$) from a noisy speech spectrum ($\left|X(\omega)\right|$), as follows.

$$\left|\hat{S}(\omega)\right| = \begin{cases} \left|X(\omega)\right| - \left|\hat{N}(\omega)\right|, & \text{if } \left|X(\omega)\right| > \left|\hat{N}(\omega)\right| \\ 0, & otherwise \end{cases} \tag{2}$$

Another enhancement scheme is called Wiener filter[16]. The Wiener filter obtains a least squares estimate of $s(t)$ under stationarity assumptions of speech and noise. The construction of Wiener filter requires an estimate of the power spectrum of the clean speech and the noise. The basic idea of Wiener filtering is to minimize the following expected value.

$$E((s(n) - \sum_{k=-\infty}^{\infty} \alpha_k \cdot x(n-k))^2) \tag{3}$$

where $s(n)$, $x(n)$, and α_k indicate the clean speech signal, the noisy speech signal, and the filter coefficient, respectively. To realize this filter in the frequency domain, it is assumed that the speech and the noise has normal distribution and do not correlate. This assumption leads the following equations.

$$E(\left|S(\omega)\right|)^2 = E(\left|X(\omega)\right|)^2 - E(\left|N(\omega)\right|)^2 \tag{4}$$

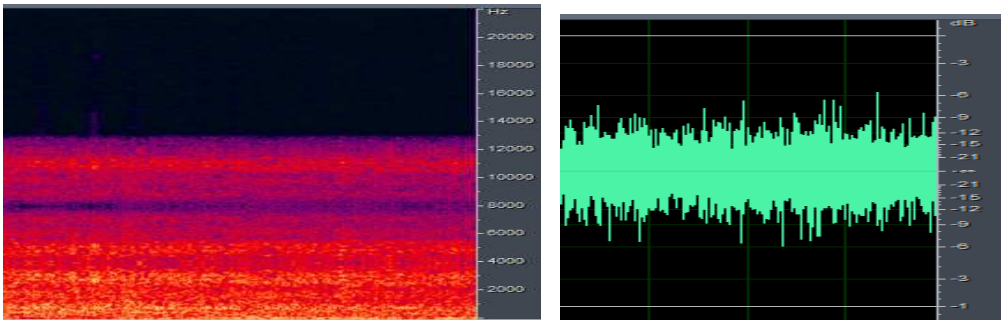
To obtain the power spectral of de-noised speech, $\left|S(\omega)\right|$, the correct estimation of a priori Signal-to-Noise Ratio (SNR) and a posteriori SNR is necessary.

3. ROBUST SPEECH AND NON-SPEECH DETECTION IN VEHICULAR ENVIRONMENTS

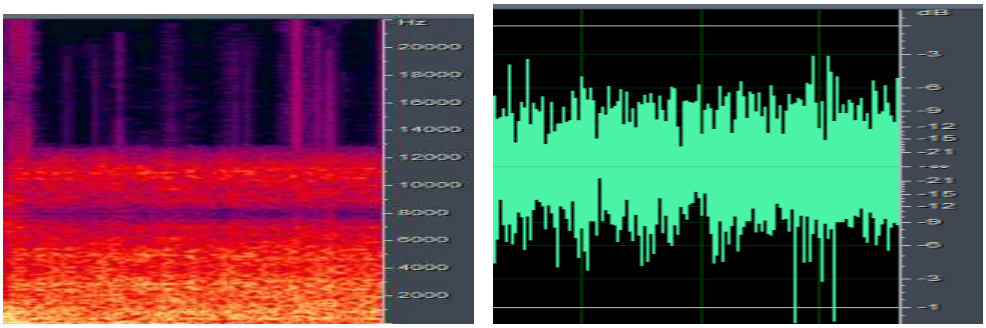
3.1 Characteristics of Noisy Speech in Vehicular Environments

There are a variety types of noise signals according to sound sources. Representative noise signals include vehicular noise like car noise and subway noise, human-oriented noise like restaurant noise and babble noise, musical noise, and so on. They are mainly categorized as stationary and non-stationary noise types. Stationary noise types indicate slowly varied signal characteristics, whereas signals of non-stationary noise types have rapidly changing property.

To observe the property of vehicular noise signals sophisticatedly, we recorded a set of noise sounds inside a car while driving the car. Figure 1 represents spectrogram and waveform plotted in frequency and time domain for the recording data. As shown in this figure, vehicular noise signals have a general property of stationary noise types, indicating slowly changing signal characteristics. The property is observed more significantly in frequency domain compared to time domain. This observation concludes that in vehicular environments, a method of reducing stationary noise signals is quite applicable.



a) Spectrogram of Recording Noise Data 1 **b) Waveform of Recording Noise Data**



c) Spectrogram of Recording Noise Data 2 **(d) Waveform of Recording Noise Data 2**

Figure 1: Spectrogram and Waveform of Noise Signals Recorded in A Car

3.2 Robust Speech and Non-Speech Detection For Vehicular Noise Reduction

Most noise reduction methods like spectral subtraction use a general property of noise signals changing more slowly compared to speech signals. And, a primary process of the methods is estimation of noise components in non-speech frames, as the estimated noise components can be used to suppress noise signals in speech frames. Thus, correct detection of non-speech regions leads to reliable noise reduction performance. In this study, we propose a method for detecting non-speech regions and estimating noise components.

In general, audio signals contaminated by stationary noise types have different spectral characteristics between non-speech regions and speech regions. In non-speech regions, spectral energy in each frequency bin is rarely varied over time because of characteristics of stationary noise. On the other hand, speech regions represent rapidly varied characteristics of spectral energy over time. In consideration of this difference, variance of spectral energy may provide a criterion of determination of speech and non-speech region. In other words, if a region indicates high variance in terms of spectral energy, the region can be regarded as a speech region. Otherwise, the region is determined to be a non-speech region.

Figure 2 summarizes the procedure of the proposed method. For a given audio region, spectral energy in each frequency bin is obtained from a sequence of audio signals. Mean and variance of spectral energy is then calculated. Next, the spectral variance of the region is compared with a pre-determined threshold, and finally, the region is determined as speech or non-speech region.

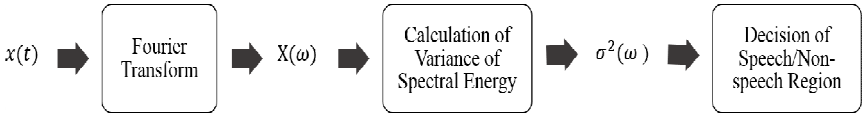


Figure 2: Procedure of the Proposed Speech/Non-Speech Decision Method

Above mechanism determining speech and non-speech regions can be described using numerical expression as follows. Note that audio signals consist of sequences of audio regions and there are k frames in a given region. Mean and variance of spectral energy of each frequency bin is calculated as follows.

$$E(\omega) = \sum_{i=0}^{k} \left| X_i(\omega) \right| \tag{5}$$

$$\sigma^2(\omega) = \frac{1}{k} \sum_{i=0}^{k} (\left| X_i(\omega) \right| - E(\omega))^2 \tag{6}$$

where $X_i(\omega)$ means spectral energy of frequency bin ω obtained from i-th frame. In non-speech regions, the variance of spectral energy in each frequency bin is expected to be a low value, conveying slowly varied characteristics. In contrast, speech-regions are expected to have a high variance. In consideration of this property, we believe that the difference between variance values in non-speech regions and those in speech regions is very significant. Hence, a pre-determined threshold defining a boundary between two different variance values can be used to decide whether a given region is a speech or non-speech region.

The proposed speech/non-speech detection method can improve the performance of noise reduction. Once a non-speech region is detected according to our mechanism, the spectral energy is regarded as the energy of noise components. Thus, the estimated spectral energy is used for suppressing noise components in subsequent speech regions.

4. EXPERIMENTAL RESULTS

4.1 Experimental Setup

In order to verify the efficiency of the proposed approach, we performed spectral analysis using audio data recorded in a car. To consider variation of noise signals, we recorded the noisy speech data in two positions (front and rear), varying engine sounds. The noisy data were recorded in three engine sound levels. Each level indicates different Signal-to-Noise Ratio (SNR) including low SNR (very noisy), moderate SNR (a little noisy), and high SNR (little noisy). For a fair comparison, the length of each recording data was fixed as 2 minutes. Table 1 summarizes positions, types and lengths of recording data used for experiments.

Table 1: Positions, Types and Lengths of Recording Data

Position	**Types and Lengths**	
	Types	***Lengths***
Front	Engine Sound Level 1 – Low SNR	2 min
	Engine Sound Level 2 – Moderate SNR	2 min
	Engine Sound Level 3 – High SNR	2 min
Rear	Engine Sound Level 1 – Low SNR	2 min
	Engine Sound Level 2 – Moderate SNR	2 min
	Engine Sound Level 3 – High SNR	2 min

A general way of evaluating the noise reduction performance is spectral analysis based on SNR. The SNR of an audio stream with T samples is estimated as follows[17].

$$SNR(dB) = 10\log_{10}\frac{\sum_{t=1}^{T}(s(t))^2}{\sum_{t=1}^{T}(n(t))^2} \tag{7}$$

where $s(t)$ represents speech signals in which noise components are reduced by noise reduction and $n(t)$ indicates noise signals. Figure 3 demonstrates an example of performance improvement based on SNR. This figure provides the change in SNR after applying noise reduction to the original noisy speech.

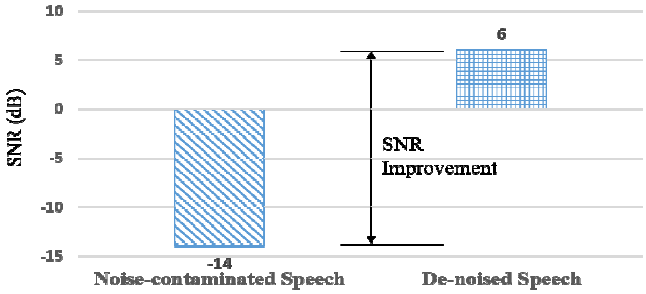


Figure 3: Example of SNR Improvement

4.2 Experimental Results

According to the SNR based evaluation scheme, we compared the SNR of the original noisy speech with that of the de-noised speech. To reduce noise components, we employed the spectral subtraction method that is the representative noise reduction method. To validate the efficiency of the proposed noise reduction approach, we investigated the SNR improvement in between the performance of the proposed approach and the conventional spectral subtraction approach. In the proposed approach, the spectral variance based VAD was applied to detect non-speech regions and estimate noise components prior to spectral subtraction, as described in Section 3.2. On the other hand, in conventional approach, the VAD was carried out using signal energy in time domain.

Table 2 and 3 summarize the results of the proposed approach and the conventional approach, respectively. The noise reduction methods successfully eliminated noise components. As shown in this table, the SNR values of the original noisy speech (denoted as 'Noise-contaminated') were improved, as compared to the SNR of the noise-reduced speech (denoted as 'De-noised'), over all positions and sound levels. In particular, the proposed noise reduction approach successfully enhanced the speech quality by reducing noise components, demonstrating significantly superior performance compared to the conventional approach. The proposed approach achieved the average SNR improvement of approximately 3dB, whereas the conventional approach showed the average improvement of 1.3dB. This result explains that the proposed VAD scheme efficiently detects non-speech regions, thus contributing to correct estimation of noise components.

Table 2: Evaluation Results of the Proposed Approach

Position	Sound Level	SNR (dB)		SNR Improvement
		Noise-Contaminated	De-Noised	
Front	Engine Sound Level 1 – Low SNR	3.98	5.36	1.38
	Engine Sound Level 2 – Moderate SNR	8.52	11.42	2.90
	Engine Sound Level 3 – High SNR	10.56	14.97	4.41
Rear	Engine Sound Level 1 – Low SNR	4.78	6.84	2.06
	Engine Sound Level 2 – Moderate SNR	9.23	12.62	3.39
	Engine Sound Level 3 – High SNR	12.30	16.34	4.04

Table 3: Evaluation Results of the Conventional Approach

Position	Sound Level	SNR (dB)		SNR Improvement
		Noise-Contaminated	De-Noised	
Front	Engine Sound Level 1 – Low SNR	3.98	4.45	0.47
	Engine Sound Level 2 – Moderate SNR	8.52	9.68	1.16
	Engine Sound Level 3 – High SNR	10.56	11.8	1.24
Rear	Engine Sound Level 1 – Low SNR	4.78	6.1	1.32
	Engine Sound Level 2 – Moderate SNR	9.23	11.21	1.98
	Engine Sound Level 3 – High SNR	12.3	13.98	1.68

CONCLUSIONS

This paper proposed an efficient noise reduction scheme to be employed for speech applications in vehicular environments. Most noise reduction methods are affected by the correctness of speech and non-speech detection, because a primary process of the methods is to estimate noise components from non-speech regions. In this paper, we proposed a robust speech and non-speech detection approach based on variance of spectral energy in frequency bin. This approach considers a general tendency of spectral energy that speech regions indicate higher spectral variance than non-speech regions do.

We evaluated the efficiency of the proposed approach using spectral subtraction. The proposed VAD method demonstrated superior noise reduction performance compared to the conventional method in terms of SNR. In future works, we will extend the noise environment including human noises like babble and musical noises.

ACKNOWLEDGMENT

This work was supported by the Agency for Defense Development(UD160054BD) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2014R1A1A2057751).

REFERENCES

1. *Kim D, Kim B. Speech recognition using Hidden Markov Models in embedded platform. Indian Journal of Science and Technology. 2015 Dec; 8(34), pp.1-4.*
2. *Park K, Park J, Bae J, Oh Y. Online speaker diarization for multimedia data retrieval on mobile devices. International Journal of Pattern Recognition and Artificial Intelligence. 2012 Dec; 26(08), pp. 1-22.*
3. *Junqua J C, Haton J P. Robustness in automatic speech recognition: Fundamentals and applications. Springer Science & Business Media. 2012.*
4. *Mattys S L, et al. Speech recognition in adverse conditions: A review. Language and Cognitive Processes. 2012; 27(7): 953-78.*
5. *Park J, Jang G, Kim J, Kim S. Acoustic interference cancellation for a voice-driven interface in smart TVs. IEEE Transactions on Consumer Electronics. 2013; 59(1):244-9.*
6. *Loizou P C. Speech enhancement: Theory and practice. New York: CRC. 2007.*
7. *Vignesh G, Sangeetha M. Surround noise cancellation and speech enhancement using sub band filtering and spectral subtraction. Indian Journal of Science and Technology. 2015 Dec; 8(33), pp.1-8.*
8. *Jang G, Park J, Kim J, Seo Y. Line spectral frequency-based noise suppression for speech-centric interface of smart devices.*

Advances in Electrical and Computer Engineering. 11(4), pp.3-8.

9. *Akagi M, Suzuki Y. A two-microphone noise reduction method in highly non-stationary multiple-noise-source environments. International Journal on IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. 2008; 91(6), pp.1337-46.*
10. *Gong Y F. Speech recognition in noisy environments: a survey. Speech Communication. 1995; 16(3), pp. 261-91.*
11. *Srinivasan S. Knowledge-based speech enhancement. Dissertation, KTH-Royal Institute of Technology, Stockholm. 2005.*
12. *Ephraim Y, Malah D. Speech enhancement using a minimum mean square error short time spectral amplitude estimator. IEEE Transactions on Acoustic Speech and Signal Processing. 1984; 32(6), pp.1109-21.*
13. *Boll S F. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust, Speech, Signal Process. 1979; 27(2), pp.113-20.*
14. *Bittu K. Mean-median based noise estimation method using spectral subtraction for speech enhancement technique. Indian Journal of Science and Technology. 2016 Sep. 9(35), pp.1-6.*
15. *Martin R. Spectral subtraction based on minimum statistics. Proc Eur Signal Process Conf; 1994. p. 1182-5.*
16. *Scalart P, Filho J, Speech enhancement based on a priori signal to noise estimation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 1996. p. 629-32.*
17. *Aoki M, Furuya K, Akitoshi K, Matsuba Y. Noise reduction method based on SAFIA for formula 1 car racing. International Workshop on Acoustic Echo and Noise Control. 2005.*